基于电源管理和工作负载整合的降低能耗解决方案

对于嵌入式开发者来说，如何在低功耗和高性能之间取得平衡，是一项艰巨而持久的工作。而电源管理和工作负载整合两种技术的诞生，正是为了帮助供应商解决这些困难与挑战。

通过基于策略性的电源管理和动态迁移来降低能耗

根据国际能源机构（IEA）最新的报告数据来看，能源消耗正在稳步上升并且在未来的一段时间仍会持续增长。该报告还预估，到 2015 年，全球的能源消耗每年将以 2.5%的速度增长，其中矿物能源消耗占据了主导地位。增长的部分主要来自于发展中国家生活方式的改变，而世界第一产业将持续为全球能源消耗的日益减少做出贡献。

业界领先的电信运营商年报显示，电信业的能源消耗持续增加，并出现在一些国家能源消耗大户的名单上。因为这些运营商持续的引入复杂的信息和通信技术，导致外围硬件设备的需求数量剧增，因此对能源的需求也随之增加，进而导致二氧化碳排放量的增加，同时能耗的成 本也随之上升。但是运营商长期的财务压力，势必要求在降低能耗支出，同时满足企业的社会责任需求和/或符合相应的法律法规。但是数据处理以及传输速率的提高，需要更多的通信设备来支持，这反过来又扩大了电信业的总体功耗。

为了获得可持续的发展，电信运营商及设备提供商开始逐渐意识到并加强电源管理技术的投入，通过重点开发能源效率计划，实现节能减排。部署于网络系统中的 AdvancedTCA（ATCA）机箱，在其整个生命周期中，大部分二氧化碳排放主要来自于机箱本身的性能需要以及冷却散热的需求。功耗则主要来自于运营阶段，在此阶段的二氧化碳排放量占整个产品生命周期总排放量的 80%左右。运营阶段中的三个层次（辅助设备、网络设备和能量转换）将消耗能量，同时也是可以管理的部分。通过对相关技术的掌握，我们可以实现能耗的管理。

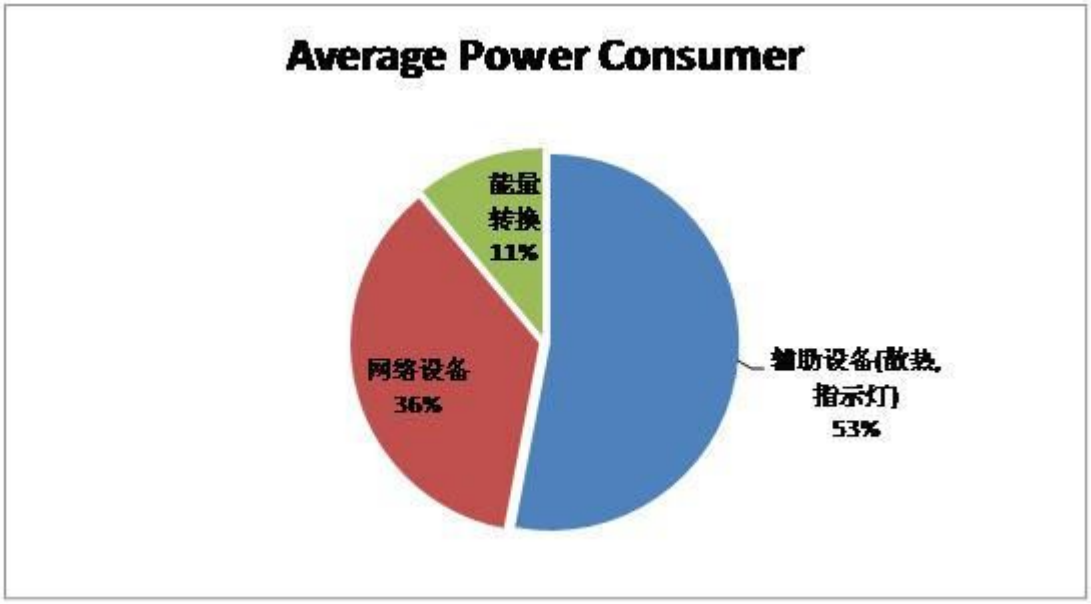


图 1. 仅有 36%的能量消耗来自于网络设备，如服务器、存储设备以及网络装置，其中大部分的能量直接转化为热能，大约只有 2.4%的能量是有效输出。如今，供应商所提供的基于 ATCA 架构的网络设备都采用了提升能源效率的解决方案，可以大幅节约辅助设备及电源转换过程中的能耗。

合理的设计对于散热管理非常重要，通过降低 CPU 的利用率，电源输出随之减少，进而降低机房内的散热需求。最终既降低了二氧化碳的排放，又减少了因散热产生的能源消耗成本。

电源管理的理念和技术

对于设备本身而言，也有一些设计理念可以用来帮助减少能耗。其中最为大家熟知的就是处理器级动态电源管理技术，这使得设备或系统可以被设置成不同的工作模式，如：性能/按需/节能/紧急。通过这项技术，可以对处理器进行动态电压调节和动态频率调节，从而进行有效的电源管理。通过动态电压调节和动态频率调节，处理器的核心电压、时钟频率或者两者都可以减小以降低能耗，同时还能满足系统的性能所需。功耗限制功能可以让系统或组件保持其能耗使用峰值在设定的数值范围内 （此数值通常根据实际的服务模式下的策略而定），如 CPU 使用率的原始数据、并发会话数量等等。

ATCA 机箱级的电源管理策略包含了用于负载整合的虚拟化动态迁移，此策略可以降低能耗和相关的成本/费用。服务器管理员可以借助动态迁移将一个正在运行的虚拟设备（VM）或应用在两个不同的物理设备间迁移，且不会断开与客户端的链接或应用。动态迁移最典型的一个应用就是云计算中的资源管理。电信运营商拥有的成千上万个虚拟设备（VM）都运行在其数据中心，为了节约能源和成本、负载均衡，这些电信运营商可以利用动态迁移对虚拟设备进行转移，而无需中断运行在这些虚拟设备中的客户应用程序。

实时迁移的配置策略可以基于能耗感知的迁移模式和/或负载调度的模式而定，这取决于首要目的是节能还是优质的服务品质。实时迁移节能的关键是有效地对服务进行打包并提供给更少的物理服务器，物理服务器数量的减少意味着对电力能源的需求就会减少，所产生的热量也随之减少，从而实现节能的最终目的。

虽然实时虚拟设备迁移具有诸多益处，如资源（CPU，内存等）的分配和能耗感知的整合，但是虚拟设备的迁移本身也需要消耗额外的能量。曾经有一篇关于虚拟设备实时迁移的性能和能量模式的文章，发表在第 20 届高性能分布式计算国际研讨会会议论文集上，该篇文章讲述了一个测试方法，用来测试实时迁移的功耗。结果显示，当部署了能耗感知以及服务器整合模型后，实时迁移所消耗的能量大幅减少。这种模式引导的决策，大幅减少了 72.9%的迁移成本，并且节能 73.6%。

配置和控制管理策略

以电信行业为例，现今的 ATCA 机箱通常包括一组高品质的电源模块以及智能风扇系统，可以用来控制温度输出和功耗。我们使用一个典型的 ATCA 机箱来做相关的测试，通过自动调整策略（根据周围的温度来决定风扇的转速），风扇（整个机箱的 1/8）的功耗可以减少 40%。

对于机箱剩余的 7/8 部分，可以通过嵌入式软件设置每个刀片上的 CPU、内存以及其他设备的频率和工作模式，从而实现动态电源管理和/或功耗限定。通过智能固件和软件层面的控制部署电源管理策略，可以大幅减少能耗。

从系统管理的角度来看，当系统的工作负载运行在满负荷水平之下时，就可以按既定策略实现动态电源管理。同时在峰值期间也可以使用动态电源管理以减少功耗。然而，当功耗（能量）节约模式启用时，处理器频率将降低，从而影响工作负载的性能和吞吐量。

功耗限定功能可以通过显示器或制动器的内部或外部处理实现。制动器可以提升处理器的电压或提升处理器/内存的频率。制动器也可以"抑制"处理器，即通过注入死循环来延迟对指令的处理。当功耗限定达到时以及限定技术启用时，工作负载的性能可能会受到影响。

嵌入式电源管理软件

电源管理软件的拓扑结构是由多个系统守护进程的组件构成，其中每个组件都会管理一个刀片，和一个客户端组件。

OFweek 通信网 www.ofweek.com

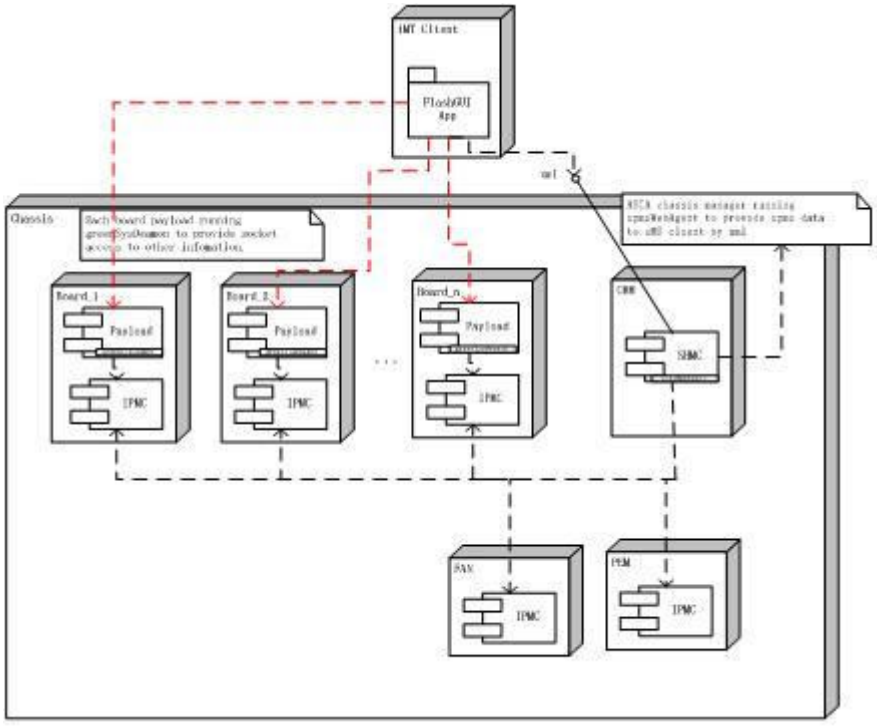


图 2：嵌入式电源管理的基本组件

客户端代表电源管理系统搜集与电源有关的数据。系统守护进程是加载在每一个刀片上的应用，扮演者电源管理模块的角色。它提供了 CPU、内存、硬盘、网络和虚拟化的工作方法以及功耗限定等功能，在满足性能需求的前提下尽量降低功耗。实际的管理端可以运行在台式机或者笔记本上，通过整合并显示输出机箱、板卡和传感器（如温度）等实际功耗的信息。

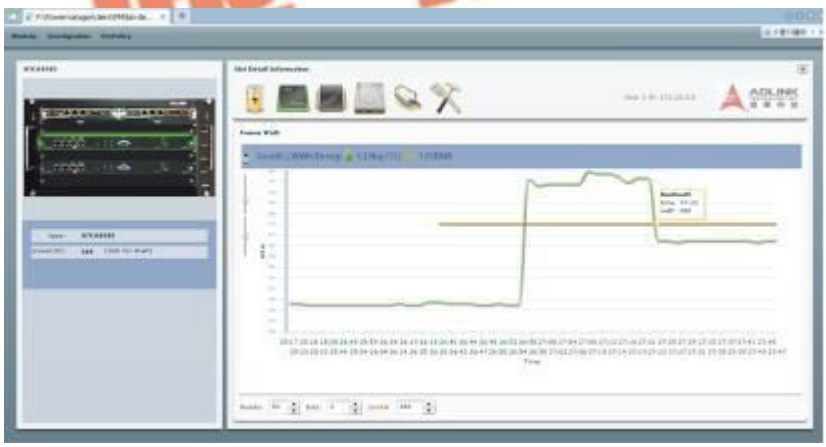
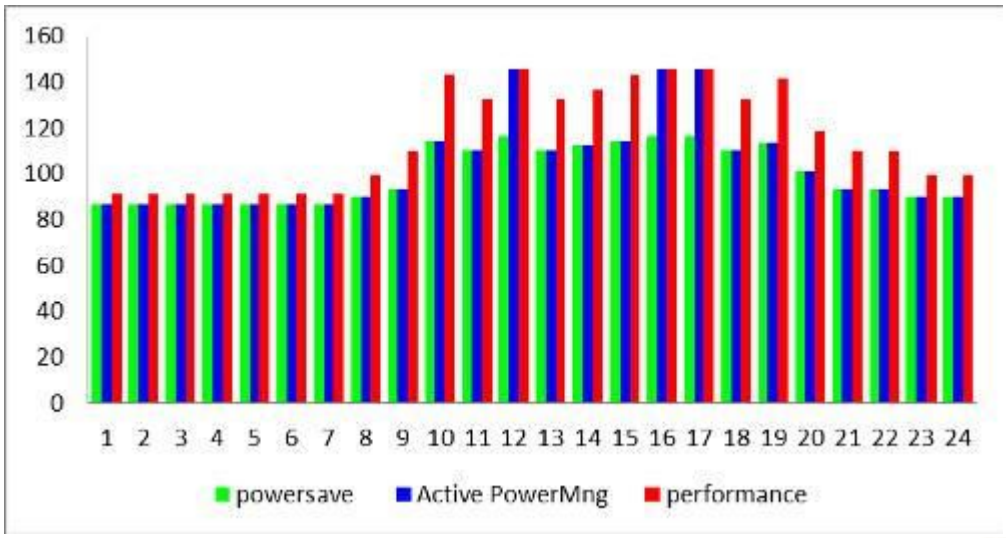


图 3：功耗限定功能实例

主动电源管理

通过策略的配置，将 ATCA 刀片上 CPU 的工作模式切换至节能或主动电源管理模式后，每个刀片的功耗相比持续运行在性能模式下减少 15%（参见图 4 和图

5）。每片板卡在加载服务的情况下可以节约 0.4KW 的功耗（参见图 5）。如果一个 14 槽的 ATCA 机框中使用了 10 个刀片，那每天节约的功耗大约 4KW。



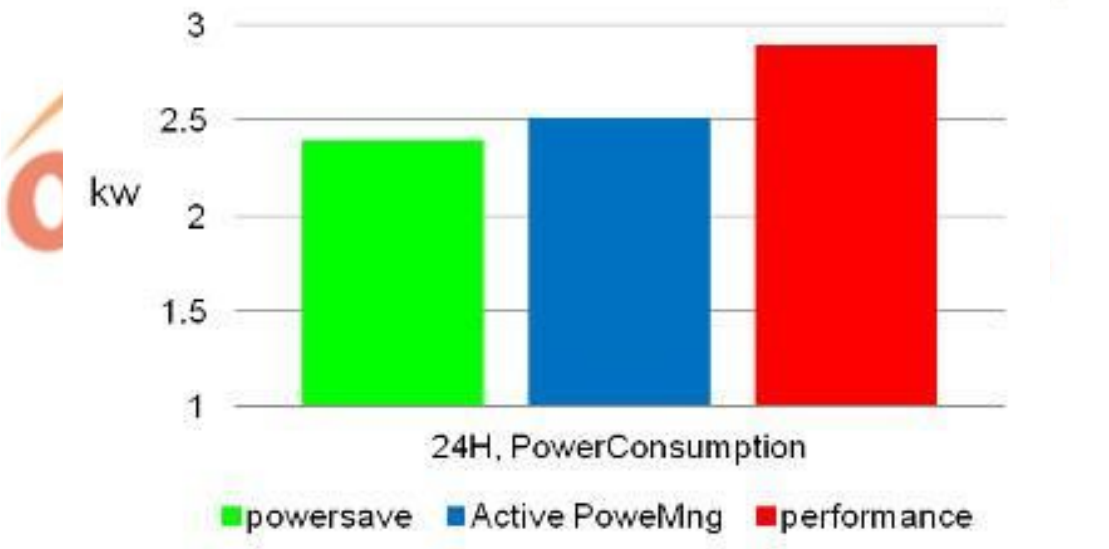


图 4 & 5：CPU 在三种独立模式下的功耗比较

动态迁移

减少功耗的另一个非常有效的方法就是只使用必要的设备来处理相关事件。利用 Erlang 概率分布算法（图表 6）可以有效检测出使用率较低的时段。

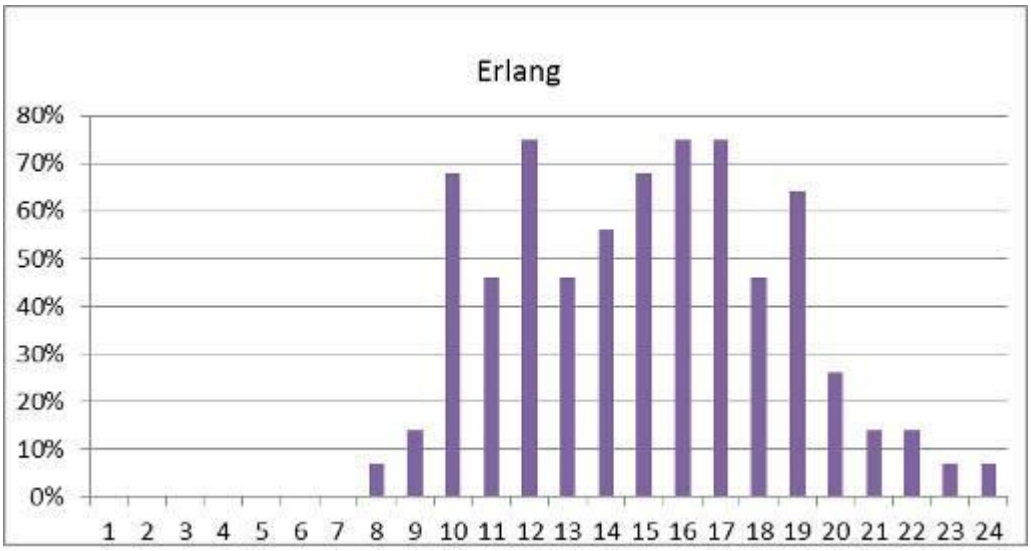


图 6：Erlang 概率分布算法在电信网络流量监测中的实例

通过上面的图表我们可以了解到，1 点至 7 点期间的 CPU 使用率最低，然而，即使运行在省电模式下，每片板卡仍然在消耗电能。在这种情况下，每片板卡在主动电源管理的策略下会消耗 90W 的功耗，峰值性能时会上升至 140W。解决的办法就是利用实时迁移策略，用最少的 CPU 刀片在处理这些工作负载，同时将节能模式下的刀片切换到睡眠模式，这样相比主动电源管理的模式可以节约超过 25%的功耗。

通过工作负载整合提升系统性能

在工作负载和 I/O 处理方面，目前的市场和技术发展趋势比较倾向采用将传统的网络架构整合到一个通用平台或模块化的组件上来，以支持多网络设备和提供不同的服务功能，如应用处理、控制处理、包处理和信号处理功能等。处理器架构以及新的软件开发工具的功能提升，让开发人员可以很容易的将工作负载整合到统一的刀片架构中，这些负载包含了应用、控制以及包处理等。通过软硬件的整合，可以大幅度提升性能，并使得刀片式服务器架构在包处理解决方案中的应用大幅增加。

为了说明工作负载整合的演变，我们设计了一系列的测试方法。这些测试方法是在单一平台中，通过将 CPU 制造商提供的 DPDK 整合到 ATCA 处理器刀片上，以此验证处理器刀片提供的性能以及整合的 IP 转发服务。比较在没有使用 Intel? DPDK 做任何优化时，采用原生 Linux（Native Linux） IP 转发时的第三层转发性能。然后，我们再分析采用 Intel? DPDK 技术之后所获得的 IP 转发性能提升的原因。

数据平面开发套件

DPDK（Data Plane Development Kit，数据平面开发套件）是一个专为 x86 架构处理器提供的轻量级运行环境。它提供了低功耗和 Run-to-Completion(RTC，

运行到完成)模式，以此最大限度的提升数据包的处理性能。而且 DPDK 还包含了优化的和高效的函数库，为用户提供丰富的选择，例如我们熟知的环境抽象层（EAL，Environment Abstraction Layer），它负责控制低级资源并提供优化的轮询模式驱动(PMD，Poll Mode Driver)，以及更高级别应用的完整 API 接口，图 7 为软件层级结构图。

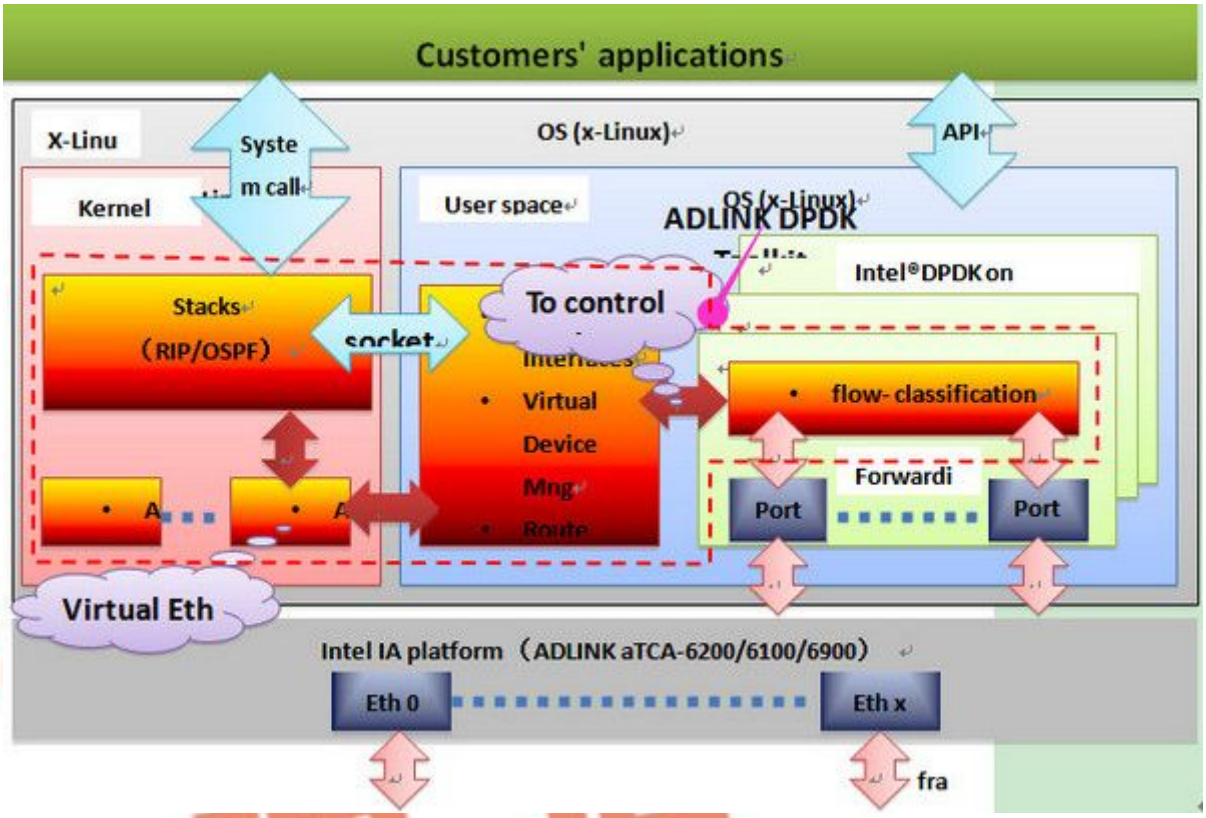


图 7： Linux 应用环境下的 EAL 和 GLIBC

测试拓扑结构

为了测量 ATCA 处理器刀片在第三层处理和转发 IP 包的速度，我们使用图 8 中所示的环境进行测试。

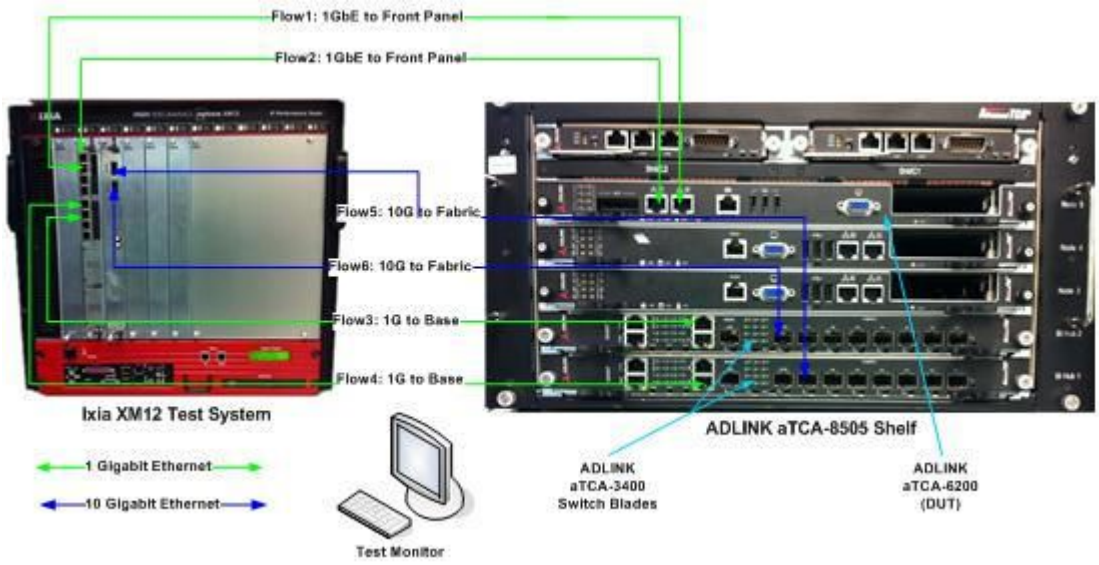


图 8：IP 转发测试环境

我们的测试使用了 ATCA 处理器刀片的 2 个 10GbE 外部接口和两个 10GbE Fabric 接口（总计 40G），通过比较使用和未使用 DPDK 的结果，我们可以得出结论：在相同的硬件平台下，使用 DPDK 后的 Linux 仅用两个 CPU 线程进行 IP 转发的性能，与原生 Linux（Native Linux）使用全部的 CPU 线程进行 IP 转发的性能相比，前者是后者的 10 倍。使用 DPDK 的平台，3 层小数据包的转发线速可以达到>70%。DPDK 中优化过的软件堆栈可以实现 10 倍性能的提升。如果在一个基于 IA 架构的刀片的控制层和数据层配备 DPDK，就可以减少一个 40G 的 NPU 刀片。通常一个 40G 的 GPU 刀片的功耗为 180W，因此通过工作负载整合可以节省 56%的能耗。

从图 9 可以看出，搭配 DPDK 后的处理器刀片的 IPv4 转发性能，可以让客户以更好的性价比成本，将包处理应用从基于硬件的网络处理器移植到基于 x86 的计算平台，同时使用同一个平台来部署不同的服务，如程序处理、控制处理和包处理服务。更多关于我们的测试过程和结果，请登录凌华科技网站 www.adlinktech.com 查询凌华科技的技术白皮书：采用 Intel? DPDK 技术的凌华科技 aTCA-6200 刀片式服务器完美实现包转发服务性能的提升。

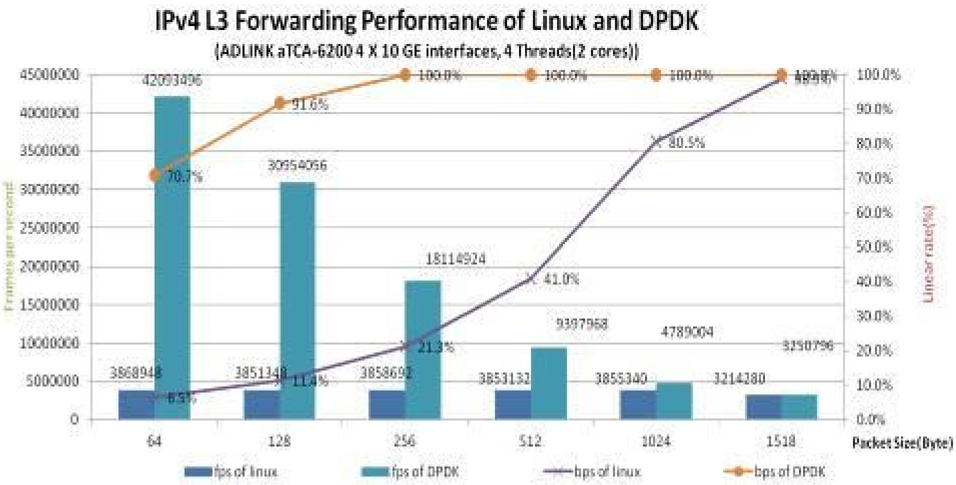


图 9：采用 4 个 10GbE 的 IP 转发性能比较

结论

目前有很多途径可以优化多板卡/多处理器系统的电源使用及效率。我们已经看到了使用嵌入式电源管理、整合嵌入式电源管理的动态迁移以及优化吞吐量的工作负载整合等方法的可能性。由于每个系统的配置和对工作负载的需求都不尽相同，因此没有一个绝对的解决办法。对于每一个方案，都需要仔细选择适合的技术和策略，以满足预期的吞吐量和功耗。

在未来，随着每个系统的功耗密度（瓦/立方英寸）的持续增加，必然对散热和运营的花费造成一定的影响，因此电源管理对于电信运营商而言将仍旧是一个需要重视的问题。

tele.ofweek.com